NBER WORKING PAPER SERIES

REGULATING TRANSFORMATIVE TECHNOLOGIES

Daron Acemoglu
Todd Lensman

Working Paper 31461
http://www.nber.org/papers/w31461

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2023

We thank Glen Weyl for several useful discussions, Joshua Gans and Chad Jones for comments, and the Hewlett Foundation and the National Science Foundation for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Regulating Transformative Technologies
Daron Acemoglu and Todd Lensman
NBER Working Paper No. 31461
July 2023
JEL No. H21,O33,O41

ABSTRACT

Transformative technologies like generative artificial intelligence promise to accelerate productivity growth across many sectors, but they also present new risks from potential misuse. We develop a multi-sector technology adoption model to study the optimal regulation of transformative technologies when society can learn about these risks over time. Socially optimal adoption is gradual and convex. If social damages are proportional to the productivity gains from the new technology, a higher growth rate leads to slower optimal adoption. Equilibrium adoption is inefficient when firms do not internalize all social damages, and sector-independent regulation is helpful but generally not sufficient to restore optimality.

Daron Acemoglu
Department of Economics, E52-446
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
daron@mit.edu

Todd Lensman
Department of Economics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
tlensman@mit.edu

1 Introduction

Recent breakneck advances in (generative) artificial intelligence have simultaneously raised hopes of productivity gains in many sectors and fears that this technology will be used for nefarious purposes, even posing an existential risk comparable to nuclear war.[1] One reaction from some experts and commentators has been a call to slow down or pause the development and adoption of AI technologies,[2] partly because a slower rollout might provide greater room for identifying danger areas and crafting appropriate regulations. There is little economic analysis of these issues, however, and it is unclear whether slowing the development and adoption of a promising, transformative technology would ever make sense.

In this paper, we develop a framework to provide a first set of insights on these questions. We consider a multi-sector economy that initially uses an old technology and can switch to a new, transformative technology. This technology is *transformative* both because it enables a higher growth rate of output and because it is general-purpose and can be adopted across all sectors of the economy. Partly because of its transformative nature, it also poses new risks. We model these by assuming that there is a positive probability of a *disaster*, meaning that the technology will turn out to have many harmful uses across a number of sectors. If a disaster is realized, some of the sectors that had previously started using the new technology may not be able to switch away from it, despite the social damages. Whether there will be a disaster or not is initially unknown, and society can learn about it over time. Critically, we also assume that the greater are the capabilities enabled by the new technology, the more damaging it will be when it is used for harmful purposes.[3]

In this environment, we study (socially) optimal and equilibrium adoption decisions. We first show that it is optimal to have a gradual adoption path, because this enables greater learning. If all sectors immediately adopted the new technology and the disaster transpired, many of them would not be able to switch back and avoid the negative social consequences. Gradual adoption instead allows society to update its knowledge and beliefs about whether this transformative technology will have socially damaging uses. Specifically, we assume that as more time passes without the disaster, the belief that there will be a disaster declines ("no news is good news"). As society thus becomes more optimistic, it is optimal to adopt the new

[1] https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[3] Both of these assumptions can be motivated with generative AI applications. For irreversibility, once large language models like ChatGPT are deployed in secondary education, it may be impossible to roll back their use, even after it becomes clear that they harm student learning. For the damages rising with productivity, many experts' fears are centered on these technologies either posing existential risks because of their capabilities or being misused, which would be more damaging when they have greater productivity (e.g., Shevlane, Farquhar, Garfinkel, Phuong, et al., 2023).

technology across a larger number of sectors. We show that under reasonable conditions this adoption path is slow and convex, accelerating only after society is fairly certain that a disaster will not occur. A simple quantitative example indicates that, for reasonable parameters on the new technology's growth advantage and disaster risk, optimal adoption can be very slow.

Perhaps surprisingly, we also show that adoption should be slower when the new technology has a higher growth rate (and damages from a disaster are large). This is for two reasons. First, since damages after a potential disaster increase with the capabilities of the new technology, a higher growth rate means that these damages also grow more quickly. Second, with a higher growth rate the effective discount rate for future output declines, so that short delays in adoption are not very consequential for discounted utility.

When compared to optimal adoption, equilibrium adoption can be inefficiently fast because private firms internalize only part of the social damages from a disaster. Even the order in which sectors adopt the new technology can differ systematically between the equilibrium and the optimum—sectors that have high social damages are not necessarily those that have high *private* damages for adopters.[4]

Finally, we discuss how regulatory schemes can help to close the gap between optimal and equilibrium adoption decisions. Pigovian taxes, use taxes, or adoption taxes that are sector-specific can fully implement the optimal adoption decisions. When sector-specific policies are not feasible, it is generally not possible to implement optimal technology choices, but regulatory actions may still be useful. In particular, it may improve welfare to prohibit use of the new technology in the sectors with the largest potential for harm until the risk of a disaster is low enough to justify broader use.

We view this paper as a first attempt to think about the consquences and regulation of transformative technologies that can be used for good or bad objectives. There are three important literatures on which we build. The first is a growing literature on economic disasters and their consequences (e.g., Rietz, 1988; Barro, 2006, 2009; Weitzman, 2009, 2011; Martin and Pindyck, 2015, 2021). This literature explores how the risk of rare economic disasters affects asset prices and cost-benefit analysis, but it typically does not focus on questions of technology adoption.

The second is a literature on technology adoption under a variety of market structures and institutional settings (e.g., Katz and Shapiro, 1986; Parente and Prescott, 1994; Foster and Rosenzweig, 1995, 2010; Acemoglu, Aghion, and Zilibotti, 2006; Acemoglu, Antràs, and Helpman, 2007; Comin and Hobijn, 2010; Comin and Mestieri, 2014). Early work touching on AI and regulation includes papers by Galasso and Luo (2018) and Agrawal, Gans, and Goldfarb

[4] For example, if AI is used to create pervasive disinformation on social media, this may be disastrous for democracy but profitable for social media platforms.

(2019), which discuss implications of privacy, trade, and liability policies for the adoption of AI. These models are similar to our framework, but they do not focus on issues of learning about social damages from new technologies.

Third, there is a nascent literature focusing on negative consequences from certain types of technologies, including environmental damages (e.g., Bovenberg and Smulders, 1995; Acemoglu, Aghion, Bursztyn, and Hemous, 2012). Most closely related to our paper are a few works that discuss the dilemma between growth and existential risk from new technologies, including AI. Jones (2023) develops a one-sector growth model in which AI can be used to raise the aggregate growth rate, but with small probability causes human extinction. Whether it is optimal to use AI depends crucially on the coefficient of relative risk aversion and whether consumption utility is bounded. Similarly, Aschenbrenner (2020) incorporates existential risk into Jones's (2016) model of growth and mortality, and argues that existential risk rises with consumption unless new mitigation technologies are developed. His model thus exhibits an "existential risk Kuznets curve" in which existential risk optimally increases until sufficient R&D resources are shifted toward mitigation. These two papers share our focus on the costs and benefits of transformative technologies, but do not address the speed of adoption across sectors and do not feature learning about risks over time. We are not aware of any other papers that incorporate these critical aspects of our work.

The rest of the paper is organized as follows. Section 2 presents our benchmark model. Sections 3 and 4 characterize optimal and equilibrium technology choices. Section 5 discusses the conditions under which optimal technology choices can be restored through regulatory taxes, and Section 6 concludes. Omitted proofs and extensions are in the Appendix.

2 Setup

We consider a continuous-time economy that linearly produces a unique final good from a continuum of sectors $i \in [0,1]$:

$$Y = \int_0^1 Y_i di.$$

A representative household has risk-neutral preferences defined over this final good and discounts the future at the rate $\rho > 0$.

Each sector can use an old technology O or a new, transformative technology N. We let $Q_j(t) > 0$ denote the time t quality of technology $j \in \{O, N\}$, and we let $x_i(t) = 1$ if sector i switches its production process to technology N, while otherwise we let $x_i(t) = 0$. Sectoral output can then be written $Y_i = (1 - x_i)Q_O + x_i \alpha_i Q_N$, where α_i designates the comparative advantage of the new technology, which may vary if the new technology is better-suited for some sectors than others. Given technology choices $x = (x_i)_{i \in [0,1]}$ and qualities $Q = (Q_O, Q_N)$, final output is

$$Y(x,Q) = \int_0^1 (1 - x_i) Q_O + x_i \alpha_i Q_N di.$$

The new technology is *transformative* in three senses. First, it is general-purpose and can be applied across all of the sectors of the economy. Second, it enables not just the production of more output, but a higher growth rate:

$$g_N > g_O \geq 0.$$

Third, because of its restructuring impact on the economy, it poses new risks. We model these by assuming that there may be a *disaster* whereby the new technology's greater productive capacity also generates negative effects. If a disaster happens, then there will be *damages* of $\delta_i Q_N > 0$ (in units of the final good) in the sectors that are using the technology. Because of possible irreversibilities, with probability $\eta_i \in (0,1)$ sector i cannot switch to technology O if it is using technology N when the disaster strikes. The realization of this reversibility event is independent across sectors. We assume that damages are proportional to Q_N because the negative effects correspond to misusing the better capabilities of the new technology.

In what follows, we reorder sectors so that δ_i is increasing and assume that i denotes the quantiles of the δ distribution, so that we can take this distribution to be uniform over some interval $[\underline{\delta}, \overline{\delta}]$. Overall damages then become

$$D(x,Q) = \left(\int_0^1 \delta_i x_i di \right) Q_N.$$

The common prior probability that there will be a disaster is $\mu(0) \in (0,1)$. If there is a disaster, we assume that the distribution of its arrival time T is exponential with rate λ. The posterior belief that there will be a disaster $\mu(t)$ evolves according to Bayes's rule:

$$\dot{\mu}(t) = -\lambda \mu(t)(1 - \mu(t)).$$

A few comments are in order. First, we model damages in each sector i by the reduced-form function $\delta_i Q_N$ to capture a broad range of potential harms from the new technology. In the context of AI, these include the spread of disinformation that harms democracy; mass unemployment; and the disruption of production in many sectors from AI-aided cyberattacks.[5]

[5] Our functional form assumptions also impose that the rate of substitution between gross consumption and

Second, as suggested above, the assumption that damages are proportional to Q_N is related to the transformative nature of this new technology. For example, if AI capabilities are used to produce disinformation, the costs will be proportional to how good these capabilities are. Third, we assume that the arrival rate of the disaster—and hence learning about the negative effects of the new technology—is independent of how many sectors switch to the new technology. This is mostly for simplicity, but is not unreasonable since many of the potential misuses of a new technology can be gradually recognized, even when this technology is not fully rolled out.[6]

3 Socially Optimal Technology Choice

In this section, we set up, solve, and provide comparative statics for the social planner's problem.

3.1 Social Planner's Problem

Given risk neutrality, the (social) planner's objective is

$$V(0) = \mathbb{E}_{\mu(0)}\left[\int_0^\infty \exp(-\rho t)\left[Y((t) - D(t)\right] dt\right], \tag{1}$$

where $Y(t)$ and $D(t)$ denote output and damages at time t and the expectation $\mathbb{E}_{\mu(0)}$ is with respect to the prior belief $\mu(0)$ over the disaster's arrival time T. To ensure that the objective is well-defined, we assume

$$\rho > g_N, \tag{2}$$

which rules out the case in which the new technology grows so quickly that discounted utility becomes infinite.

It is more convenient to work with the recursive formulation of (1), which has three state variables: the posterior belief of disaster, μ; the time-varying qualities of the old and new technologies, Q; and, after the disaster (if any), the set of sectors that were already using the new technology and for which this use is irreversible. We track these sectors using the vector $\bar{x} = (\bar{x}_i)_{i\in[0,1]}$, where $\bar{x}_i = 1$ if sector i uses technology N irreversibly and $\bar{x}_i = 0$ otherwise. Let $V(\mu, Q)$ denote pre-disaster household welfare, and let $W(\bar{x}, Q)$ denote post-disaster welfare.

damages in utility is constant and equal to one. Jones (2023) points out that this may not hold in the case of existential risk and explores the implications for optimal use of a life-threatening new technology.

[6] Alternative assumptions are discussed in Section 6.

Then the Hamilton-Jacobi-Bellman (HJB) equations for the planner are

$$\rho V(\mu,Q) = \max_{x_i\in\{0,1\}} \{Y(x,Q) + \mu\lambda(\mathbb{E}[W(\bar{x},Q)|\,x] - V(\mu,Q))\} + \dot{V}(\mu,Q), \tag{3}$$

$$\rho W(\bar{x},Q) = \max_{x_i\in\{\bar{x}_i,1\}} \{Y(x,Q) - D(x,Q)\} + \dot{W}(x,Q). \tag{4}$$

Equation (4) imposes that x_i cannot be less than $\bar{x}_i$, since if $\bar{x}_i = 1$ sector i's use of the new technology has turned out to be irreversible. Given this, V depends on the conditional expectation of welfare after a disaster given the current technology choices x, denoted by $\mathbb{E}[W(\bar{x},Q)|\,x]$.[7] In (3) we also use the fact that the arrival rate of the disaster, given the posterior μ, is $\mu\lambda$.

To characterize the planner's technology choices, suppose first that the disaster has occurred. The planner's problem in (4) is linear, so the solution is

$$x_i = \begin{cases} 1 & \text{if} \quad \bar{x}_i = 1 \quad \text{or} \quad (\alpha_i - \delta_i)Q_N > Q_O, \\ 0 & \text{else.} \end{cases}$$

This expression imposes, without loss of generality, that the planner sticks with the old technology if indifferent. It also incorporates the fact that the planner is constrained to choose $x_i = 1$ if $\bar{x}_i = 1$. Even when unconstrained, it may be optimal to set $x_i = 1$ if the output produced by technology N exceeds its damages plus the output that can be produced by technology O. In the remainder of the text we assume that damages are sufficiently large that, whenever possible, the planner chooses technology O after a disaster:

$$\alpha_i \leq \delta_i. \tag{5}$$

This enables us to focus on the most interesting case where damages exceed the benefits of the new technology. The general case is studied in Appendix B.

Integrating the HJB equation (4) and taking expectations with respect to $\bar{x}$, we have

$$\mathbb{E}[W(\bar{x},Q)|\,x] = \int_0^1 \left[\left(1 - x_i\eta_i\right)\frac{1}{\rho - g_O}Q_O + x_i\eta_i\frac{\alpha_i - \delta_i}{\rho - g_N}Q_N\right]di.$$

Then, before the disaster, it is optimal from (3) to use technology N in sector i iff

$$\alpha_i Q_N - Q_O > \mu\lambda\eta_i\left[\frac{1}{\rho - g_O}Q_O - \frac{\alpha_i - \delta_i}{\rho - g_N}Q_N\right]. \tag{6}$$

Intuitively, the left-hand side is the flow gain from using technology N in sector i, while the

[7] To determine this conditional expectation, we use $\mathbb{P}(\bar{x}_i = 1|\,x_i = 1) = \eta_i$ and $\mathbb{P}(\bar{x}_i = 1|\,x_i = 0) = 0$.

right-hand side is the expected loss due to the disaster, including both the discounted value of lost output and the irreversible damages. These losses are multiplied by the posterior arrival rate of the disaster $\mu\lambda$ and the probability of irreversibility η_i. Since μ is decreasing and Q_N/Q_O is nondecreasing, for any initial state $(\mu(0), Q(0))$ there exists a time $t_i < \infty$ such that technology O is used in sector before t_i and technology N is used thereafter.

3.2 Socially Optimal Technology Adoption

To determine how (socially) optimal use of technology N changes over time, denote the fraction of sectors that use technology N, or total *adoption*, by

$$X(\mu, q) = \int_0^1 x_i(\mu, q)\, di,$$

where $q = \log(Q_N/Q_O)$ is the *quality gap* between the technologies, and $x_i(\mu, q) = 1$ iff it is optimal to use technology N in sector i in state (μ, q). For simplicity, also assume that α_i and η_i are constant across sectors, and denote the cumulative distribution function of the uniform distribution over $[\underline{\delta}, \bar{\delta}]$ by F. These assumptions imply that there exists a *damage threshold* $L(\mu, q)$ such that it is optimal to adopt the new technology in sector i iff $\delta_i < L(\mu, q)$. Total adoption of the new technology is then just the fraction of sectors below the damage threshold:

$$X(\mu, q) = F(L(\mu, q)).$$

The following proposition is immediate from (6), and we omit its proof:

Proposition 1. *It is socially optimal to use technology N in sector i iff $\delta_i < L(\mu, q)$, where*

$$\frac{L(\mu, q) - \alpha}{\rho - g_N} = \frac{\alpha - \exp(-q)}{\mu\lambda\eta} - \frac{\exp(-q)}{\rho - g_O}. \tag{7}$$

$L(\mu, q)$ (and thus $X(\mu, q)$) is strictly increasing in α and q; strictly decreasing in g_O, λ, and μ; and strictly decreasing in g_N, provided that $L(\mu, q) > \alpha$.

In light of our assumption (5), the condition $L(\mu, q) > \alpha$ is satisfied as soon as there is any adoption of the new technology. Proposition 1 then implies that when the new technology enables *faster growth*, its adoption should be *slower*. This is because of a *precautionary channel*—even though the planner is risk neutral, she would like to avoid the risk of irreversible damages from the new technology, and this introduces a precautionary motive. The faster the new technology grows, the greater are its potential damages as well, and this strengthens the precautionary motive.[8]

The comparative statics in Proposition 1 are partial, because they hold the state (μ, q) fixed. Full comparative statics must account for how parameter changes affect the evolution of the state $(\mu(t), q(t))$. The belief $\mu(t)$ does not depend on the growth rates g_O and g_N, but the quality gap $q(t)$ does, since $q(t) = q(0) + (g_N - g_O)t$. Because the damage threshold $L(\mu, q)$ is increasing in the quality gap, any change in the growth rates affects adoption at each time $t > 0$ through both the direct effects described in Proposition 1 and the indirect effects through changes in the quality gap $q(t)$. The next proposition characterizes the total effect of a change in technology growth rates.

Proposition 2. *For all $t > 0$:*

1. *$X(\mu(t), q(t))$ is decreasing in g_O.*
2. *There exists an earliest time $\bar{t} < \infty$ such that $X(\mu(t), q(t))$ is decreasing in g_N if $t > \bar{t}$. The time $\bar{t}$ is decreasing in g_N.*
3. *Adoption falls to zero as g_N approaches ρ, i.e., $\lim_{g_N \uparrow \rho} X(\mu(t), q(t)) = 0$.*

The proof of this proposition and other results in the text are presented in Appendix A, unless otherwise stated.

The first part of Proposition 2 establishes that the comparative static for g_O from Proposition 1 generalizes in the presence of the indirect effects through $q(t)$—the quality gap $q(t)$ is declining in g_O, reinforcing the direct effect. The second part shows that the new technology's growth rate has more nuanced implications: Adoption is not always decreasing in g_N, but is after some critical time $\bar{t}$, and this time itself is a decreasing function of g_N. In this case, the precautionary channel highlighted above must compete with the fact that the quality gap $q(t)$ is increasing in g_N, but this indirect effect can dominate only at short time horizons.

The third part of proposition establishes that as g_N increases towards the discount rate, adoption almost stops. This might appear paradoxical initially, but it is also intuitive. When g_N is approximately equal to ρ, the benefits from the new technology are very high, leading to nearly infinite discounted utility provided no disaster arrives. Delay in initiating the adoption of the technology has little effect on these benefits. However, a disaster will have huge negative consequences, and avoiding this disaster now takes precedence.

[8] This holds because, under (5), post-disaster net output is decreasing in Q_N in each sector using the new technology. In Appendix B, we show that when this assumption is relaxed, so that post-disaster net output can be increasing in Q_N for some sectors i with $\delta_i < \alpha$, the damage threshold $L(\mu, Q)$ and adoption $X(\mu, Q)$ may be increasing in g_N.

The next proposition further characterizes the shape of the adoption curve. Since F is uniform, $\dot{X}(\mu,q) = f\dot{L}(\mu,q)$, where f is the constant density of F. Hence, the *curvature* of technology adoption can be written

$$\frac{\ddot{X}(\mu,q)}{\dot{X}(\mu,q)} = \frac{\ddot{L}(\mu,q)}{\dot{L}(\mu,q)}.$$

We therefore have:

Proposition 3.

1. *$\dot{L}(\mu,q)$ is strictly decreasing in g_O, and it is strictly decreasing in g_N iff the quality gap is sufficiently large, i.e.,*

$$\alpha\exp(q)-1 > \frac{(\rho-g_N)-(g_N-g_O)}{1-\mu}\left(\frac{1}{\lambda}+\frac{\mu\eta}{\rho-g_O}\right).$$

2. *There exists a positive constant $G(\mu,q)$ such that if $\alpha\exp(q) > 1$, $\ddot{L}(\mu,q)$ is positive iff $g_N - g_O > G(\mu,q)$. $G(\mu,q)$ is independent of g_N and increases to infinity over time.*

The intuition for the first part of the proposition is the same as for Proposition 2: The damage threshold increases as the posterior belief μ falls and the quality gap q grows. When technology O grows more quickly, it slows the rate of increase in the quality gap and raises the opportunity cost of using technology N after the disaster for a fixed quality gap. As a result, the damage threshold grows less quickly in each state. When technology N grows more quickly, it raises both the rate of increase in the quality gap and the net output losses from technology N after the disaster. The latter effect dominates when the quality gap is sufficiently large because additional improvements in technology N relative to O have only a negligible impact on the planner's technology choice.[9]

The second part of the proposition proves that when the new technology's growth advantage is sufficiently large, its adoption will eventually have a major convex segment in which its adoption accelerates. This result holds despite the fact that the learning rate $|\dot{\mu}|$ is declining at a greater than exponential rate when $\mu < \frac{1}{2}$ (in particular, $\frac{d}{dt}|\dot{\mu}| = -\lambda|\dot{\mu}|(1-2\mu)$). This is because expected damages from technology N in sector i are proportional to the posterior μ, and as μ declines, larger increases in the damage threshold $L(\mu,q)$ are needed to balance the expected damages and benefits in the "marginal" sector.[10]

[9] The latter effect also dominates regardless of the quality gap whenever $L(\mu,z) > 0$ and $g_N - g_O \geq \rho - g_N$.

[10] In Appendix C, we verify this intuition by showing that learning dynamics favor *concave* adoption when sectors are heterogeneous according to α_i instead of δ_i.

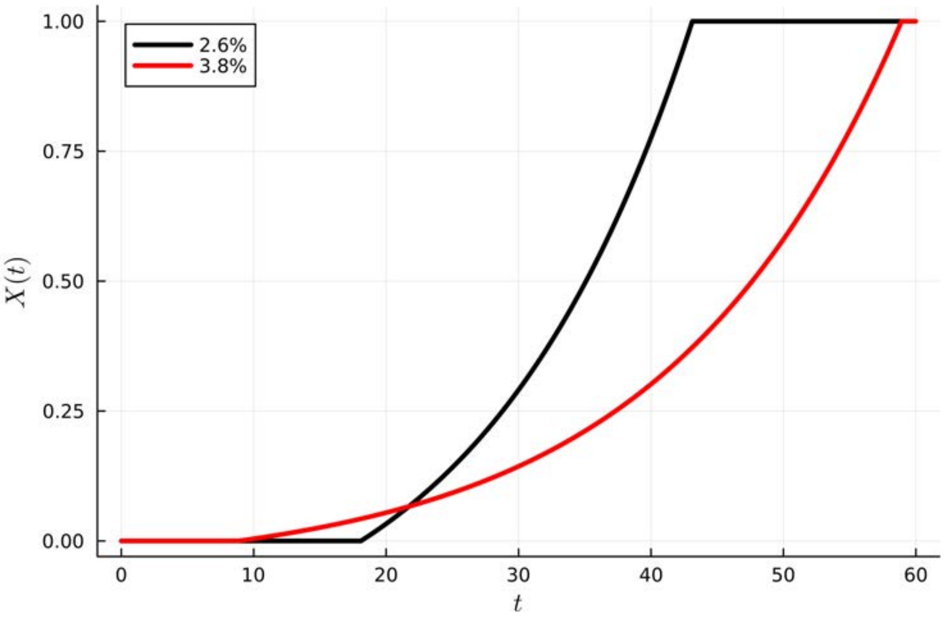

Figure 1: Adoption curves $X(t) \equiv X(\mu(t), q(t))$ for different values of g_N. The remaining parameter values are $\rho = 0.04$, $\lambda = 0.05$, $\eta = 0.5$, $\alpha = 1$, $g_O = 0.02$, $\underline{\delta} = 1$, and $\bar{\delta} = 5$. The initial state is $\mu(0) = 0.2$ and $q(0) = 0$.

We end this section by depicting the time path of adoption in a couple of parameterized cases in Figure 1. We set $g_O = 2\%$ in line with trend GDP growth in developed economies and $\rho = 0.04$ to produce a risk-free interest rate of 4%. We choose two values for g_N based on Chui, Roberts, Yee, Hazan, et al. (2023), who forecast an increase in the growth rate of 0.6-3.6% in the United States between 2023 and 2040 from AI and other automation technologies. We take the lower end of this range, $g_N - g_O = 0.6\%$, and a higher but still conservative estimate from the middle of the range: $g_N - g_O = 1.8\%$ (while still satisfying (2)). We take the two technologies to have the same quality in year $t = 0$, thus $q(0) = 0$. We suppose that the range of damages is from one to five times gross sectoral output ($\underline{\delta} = 1$, $\bar{\delta} = 5$), and we set $\eta = 0.5$ so that half of all sectors using the new technology cannot switch back after a disaster. We set the expected arrival time of a disaster (if one exists) to be 20 years, which gives $\lambda = 0.05$. Finally, a recent survey of AI experts reports a median estimate of existential risk of about 10%,[11] and since we are interested in non-existential misuses of AI as well, we choose the initial disaster probability to be twice as large, $\mu(0) = 20\%$. Figure 1 shows that optimal adoption is slow, taking about 40 years until full adoption when $g_N = 2.6\%$ and almost 60 years when $g_N = 3.8\%$.

In summary, this section has established that the optimal adoption of a new, transformative

[11] https://aiimpacts.org/2022-expert-survey-on-progress-in-ai

technology should be slow and gradual, particularly when its superior capabilities also make its potential damages greater and there is learning about the likelihood of misuse (a “disaster”). Notably, a higher growth rate for a transformative technology can lead to slower optimal adoption.

4 Equilibrium Technology Choice

We now turn to a characterization of equilibrium technology adoption, assuming that private firms do not internalize all social damages after a disaster.

4.1 The Firm’s Problem

Suppose now that in each sector, the choice of technology is made by a private (representative) firm that seeks to maximize expected discounted profits. To simplify, we assume that the firm in sector i appropriates all output of its intermediate as profits, but only internalizes *private damages* $\gamma_i \leq \delta_i$.

The firm’s profit maximization problem can be formulated recursively in the same way as the planner’s problem in the previous section. The state variables before the disaster are again μ and Q, and the state variables relevant for firm i after the disaster are $\bar{x}_i$ and Q. Let $\Pi_i(\mu, Q)$ denote the firm’s pre-disaster value, $\Phi_i(\bar{x}_i, Q)$ the firm’s post-disaster value, and $Y_i(x_i, Q)$ its (gross) output. The HJB equations for the firm are

$$\rho\Pi_i(\mu,Q) = \max_{x_i\in\{0,1\}} \left\{Y_i(x_i,Q) + \mu\lambda\left(\mathbb{E}\left[\Phi_i(\bar{x}_i,Q)\middle| x_i\right] - \Pi_i(\mu,Q)\right)\right\} + \dot{\Pi}_i(\mu,Q), \tag{8}$$

$$\rho\Phi_i(\bar{x}_i,Q) = \max_{x_i\in\{\bar{x}_i,1\}} \left\{Y(x_i,Q) - x_i\gamma_i Q_N\right\} + \dot{\Phi}_i(\bar{x}_i,Q). \tag{9}$$

These value functions differ from the planner’s, (3) and (4), because the firm internalizes only a fraction γ_i/δ_i of the flow damages from technology N.

We assume that private damages are also sufficiently large that firm i will always choose technology O after the disaster if possible:[12]

$$\alpha_i \leq \gamma_i. \tag{10}$$

[12] Without this assumption, the equilibrium would be even more inefficient, as firms will continue to use the new technology in some (reversible) sectors even after a disaster.

Similar to the planner's problem, it is privately optimal for firm i to use technology N iff

$$\alpha_i Q_N - Q_O > \mu\lambda\eta_i\left[\frac{1}{\rho - g_O}Q_O - \frac{\alpha_i - \gamma_i}{\rho - g_N}Q_N\right].$$

The only difference between this condition and the planner's optimality condition (6) is that private damages γ_i appear instead of social damages δ_i on the right-hand side. Firm i internalizes fewer damages from irreversible use of technology N after the disaster, and as a result it begins using technology N earlier than the planner before the disaster.

4.2 Equilibrium Technology Adoption

We denote total equilibrium adoption by

$$\tilde{X}(\mu,q) = \int_0^1 \tilde{x}_i(\mu,q)\,di,$$

where $\tilde{x}_i(\mu,q) = 1$ iff firm i uses technology N in state (μ,q). Again assuming that α_i and η_i are constant across sectors, it is immediate to see that firm i will adopt the new technology iff *private* damages are lower than the damage threshold, $\gamma_i < L(\mu,q)$. Equilibrium adoption is then

$$\tilde{X}(\mu,q) = F_\gamma(L(\mu,q)),$$

where F_γ is the cumulative density function of γ_i.

This characterization implies that all comparative statics results from Section 3.2 also apply to equilibrium adoption. In particular, the results in Propositions 1 and 3 concern only the damage threshold $L(\mu,q)$ and apply exactly as stated, while Proposition 2 applies after replacing $X(\mu,q)$ with $\tilde{X}(\mu,q)$ (and so we omit their proofs):

Proposition 4. *For all $t > 0$:*

1. *$\tilde{X}(\mu(t),q(t))$ is decreasing in g_O.*
2. *There exists an earliest time $\tilde{t} < \infty$ such that $\tilde{X}(\mu(t),q(t))$ is decreasing in g_N if $t > \tilde{t}$. The time $\tilde{t}$ is decreasing in g_N.*
3. *Adoption falls to zero as g_N increases to ρ: $\lim_{g_N \uparrow \rho}\tilde{X}(\mu(t),q(t)) = 0$.*

In the remainder of this section, we seek to understand how the optimal and equilibrium adoption curves differ when private damages γ_i diverge from social damages δ_i. First observe that even similar adoption curves do not imply that the equilibrium is optimal, because the order in which sectors adopt the new technology matters. For example, private and social damages may be *negatively affiliated*, meaning that high social damage sectors have low private damages. In this case, the order in which the new technology spreads in equilibrium is exactly the opposite of the optimal order.

Even when equilibrium and optimal orders of adoption coincide, the equilibrium can be inefficient. To see this, suppose that social and private damages are *positively affiliated*, so that there exists a non-negative and strictly increasing function κ with $\gamma_i = \kappa(\delta_i) \leq \delta_i$. Then we can write equilibrium adoption in terms of the (uniform) distribution of social damages F:

$$\tilde{X}(\mu, q) = F\left(\kappa^{-1}(L(\mu, q))\right).$$

This equation implies that the equilibrium adoption curve $\tilde{X}(\mu(t), q(t))$ is a distorted version of the optimal adoption curve, with an *equilibrium damage threshold* $\tilde{L}(\mu, q) = \kappa^{-1}(L(\mu, q))$. In this case, knowing how the equilibrium and social damage thresholds differ is sufficient to fully characterize the inefficiencies in equilibrium adoption. The next proposition determines how the level, rate of change, and curvature of the equilibrium damage threshold $\tilde{L}(\mu, q)$ differ from its social counterpart $L(\mu, q)$.

Proposition 5.

1. *The equilibrium damage threshold is always greater than the social damage threshold:* $\tilde{L}(\mu, q) \geq L(\mu, q)$.

2. *The equilibrium damage threshold increases more quickly than the social damage threshold, provided that* $\kappa'\left(\tilde{L}(\mu, q)\right) < 1$*:*

$$\dot{\tilde{L}}(\mu, q) = \frac{\dot{L}(\mu, q)}{\kappa'\left(\tilde{L}(\mu, q)\right)}.$$

3. *The equilibrium damage threshold is more convex than the social damage threshold when κ is locally concave, i.e.,*

$$\frac{\ddot{\tilde{L}}(\mu, q)}{\dot{\tilde{L}}(\mu, q)} = \frac{\ddot{L}(\mu, q)}{\dot{L}(\mu, q)} - \frac{\kappa''\left(\tilde{L}(\mu, q)\right)}{\kappa'\left(\tilde{L}(\mu, q)\right)}\dot{L}(\mu, q).$$

These results follow from the definition of the equilibrium damage threshold $\tilde{L}(\mu, q)$. We illustrate them in Figure 2 by depicting socially optimal and equilibrium adoption curves for the benchmark parameterizations in Figure 1 and a concave affiliation function κ. We see that the equilibrium damage threshold is always greater than its social counterpart, and it increases

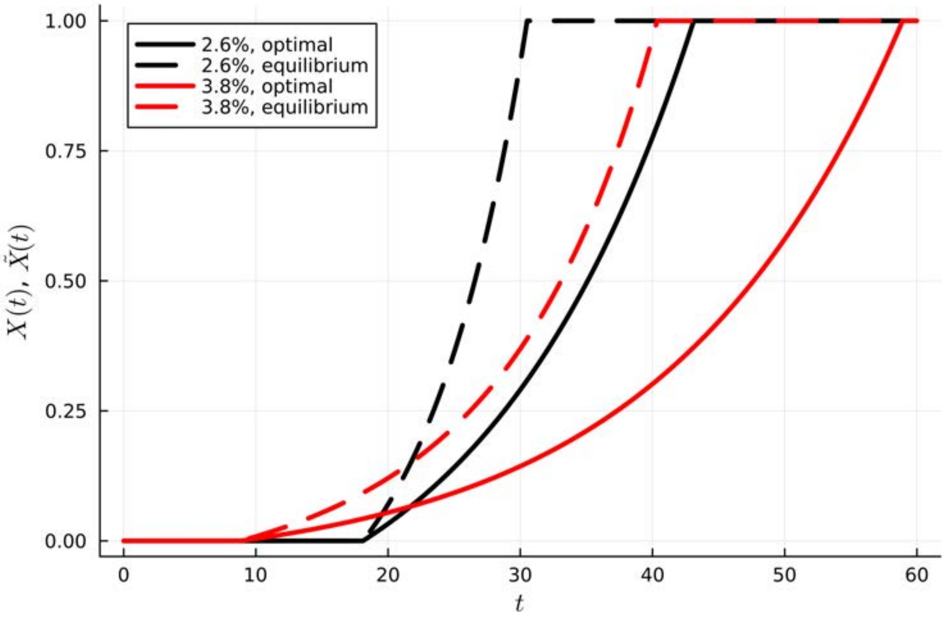

Figure 2: Socially optimal and equilibrium adoption curves, $X(t)$ and $\tilde{X}(t) \equiv \tilde{X}(\mu(t), q(t))$. The calibration is the same as in Figure 1. The affiliation function is $\kappa(\delta) = \delta^{1/2}$.

more quickly (for the marginal sectors where $\kappa'\left(\tilde{L}(\mu, q)\right) < 1$). Consequently, equilibrium adoption is inefficiently rapid and accelerates when there are high social damages.

In summary, equilibrium adoption of transformative technologies is determined by the same forces that shape optimal adoption. However, because firms are motivated by higher productivity and discouraged only by private damages, equilibrium adoption is generally suboptimal: Firms do not fully internalize social damages from potential disasters, so equilibrium adoption is typically too high and rises too quickly, and the order in which sectors adopt the new technology may differ from the optimal one.

5 Regulating Technology Choice

Since equilibrium adoption is potentially inefficient, a natural question is whether government regulation can close the gap between equilibrium and optimal adoption decisions. Throughout this section, we simplify the analysis by focusing on *ex ante* regulations.[13]

Socially optimal and equilibrium technology choices differ because the planner and private firms internalize different damages after the disaster and hence different *expected* damages

[13] We ignore *ex post* (“Pigovian”) taxes both because their analysis is essentially identical to our characterization of use taxes, and also because they may not be credible as they do not affect technology choice after the disaster—the private sector already stops using the new technology whenever possible.

before the disaster. A straightforward way to correct firms' incentives is through a *use tax* that raises firms' costs of using the new technology *before the disaster*.[14] If sector-specific taxes are feasible, then the tax that implements the optimal technology choice for sector i is equal to the difference between expected discounted social and private damages:

$$\tau_i(\mu, Q_N) = \mu\lambda\eta_i \frac{\delta_i - \gamma_i}{\rho - g_N} Q_N. \tag{11}$$

The next proposition notes several properties of these optimal taxes.

Proposition 6. *The optimal use tax $\tau_i(\mu, Q_N)$ is larger in sectors with a larger probability of irreversibility η_i and a larger difference between social and private damages $\delta_i - \gamma_i$. It is log-concave in time and limits to zero as $t \to \infty$ iff $\lambda > g_N$.*

The cross-sector comparative statics follow immediately from equation (11). Differentiating it with respect to time yields

$$\frac{\dot{\tau}_i(\mu, Q_N)}{\tau_i(\mu, Q_N)} = \frac{\dot{\mu}}{\mu} + \frac{\dot{Q}_N}{Q_N} = -\lambda(1-\mu) + g_N.$$

Since μ declines before the disaster, we observe that $\tau_i(\mu(t), Q_N(t))$ is log-concave. The assumption that the social and private damages from a disaster are increasing in Q_N provides a force for the tax to increase over time, while growing optimism about the absence of a disaster pushes taxes lower. The tax is eventually decreasing to zero iff learning about the disaster risk is sufficiently fast, $\lambda > g_N$.

Sector-specific taxes require a planner to have very detailed information about damages and may generally be difficult to implement. Even in the benchmark case in which α_i and η_i are constant across sectors, the next proposition shows that a sector-*independent* tax scheme cannot correct inefficient equilibrium adoption unless social and private damages are positively affiliated.

Proposition 7. *Given any sector-independent use tax $\tau(\mu, Q)$, firm i begins using technology N earlier than firm j iff $\gamma_i \leq \gamma_j$. Socially optimal technology choices can be implemented for any initial state $(\mu(0), Q(0))$ iff social and private damages are positively affiliated. In this case, the following tax is optimal:*

$$\tau(\mu, Q) = \mu\lambda\eta \frac{L(\mu, q) - \kappa(L(\mu, q))}{\rho - g_N} Q_N. \tag{12}$$

[14] Naturally, *adoption taxes* that are paid when new technologies are first introduced are equivalent to these use taxes.

When private and social damages are not positively affiliated, the power of sector-independent tax and regulatory schemes declines even further, because the order in which different sectors adopt the new technology cannot be aligned with the social optimum. In this case, a different policy that we refer to as a *regulatory sandbox* may be more effective. Under this policy, sectors with social damages below a threshold $\hat{\delta}$ ("inside the sandbox") can choose their technology freely, while sectors above the threshold are restricted from using the new technology until time $\hat{T}$. This policy allows the planner to ensure that sectors with high social damages adopt only after the new technology is established to be relatively safe. The next proposition demonstrates that the sandbox policy can improve upon the laissez-faire equilibrium.

Proposition 8. *Suppose α_i and η_i are constant and $\gamma_i < \delta_i$ across all sectors i. Then there exists a sandbox policy $\left(\hat{\delta}, \hat{T}\right)$ that strictly improves upon the laissez-faire equilibrium.*

In Appendix D, we provide additional details about optimal regulatory sandboxes and compare them to sector-independent taxes. In general, combining a sector-independent tax with a regulatory sandbox is better than either policy alone: A sector-independent tax can differentially delay adoption for sectors with varying private damages γ_i, but it cannot alter the order of adoption. In contrast, a regulatory sandbox can alter the order by delaying adoption for sectors with high social damages.

6 Concluding Remarks

Advances in generative AI technologies, such as GPT-4 and other large language models, have both raised hopes of more rapid growth thanks to the rollout of these technologies and concerns about misuses and unforeseen negative consequences from their new capabilities. Despite a multifaceted public discussion about their regulation, there are currently no economic models of the regulation of transformative technologies. This paper has taken a first step in building such a model to provide insights for this debate.

We consider the adoption decision of a new, transformative technology that can increase productivity growth across all sectors of the economy but also raises risks of misuse, which we model as the stochastic arrival of a "disaster". If a disaster occurs, some of the sectors that started using the new technology may be unable to switch back to the old, safe technology and generate social damages. We assume that the likelihood of a disaster is unknown and society gradually learns about whether such a disaster will occur. We show that adoption should be slow and follow a convex path, initially growing slowly before accelerating later. This slow adoption is motivated by social learning about the likelihood of a disaster—as the posterior probability of a disaster declines over time, adoption increases. Most surprisingly, a faster growth rate of the new technology should lead to slower adoption. This is because of a precautionary channel: Despite the fact that the planner is risk neutral, irreversible damages imply that it is optimal to wait and learn about the likelihood of a disaster. These irreversible damages are greater when the new technology has a higher growth rate, strengthening the precautionary motive. Finally, if private firms internalize only part of the social damages from transformative technologies, equilibrium adoption tends to be too fast and necessitates regulatory policies, some of which we characterized.

There are many interesting areas left for future work. First, we assumed, both as a natural benchmark and for tractability, that the rate at which society learns about the likelihood of a disaster is independent of which sectors have adopted the technology. In practice, early adoption may increase risks or may facilitate learning. In addition, there may be sector-specific learning about "safe use" of the new technology. These considerations may motivate "experimentation" by adopting the technology in a few sectors or trying different uses in some areas. An analysis of these types of experimentation is an interesting area for future work. Second, many of the misuses of new AI technologies depend on market structure and other aspects of regulation (e.g., concerning disinformation, discrimination, or privacy), and it would be interesting to explore how these affect optimal and equilibrium adoption. Third, we simplified the analysis by assuming risk neutrality. As explored in Jones (2023), the extent of risk aversion in the social welfare function has a first-order effect on the trade-off between higher growth and the likelihood of a disaster, and these can be incorporated in future analyses of learning about misuses of new transformative technologies. Fourth, we also abstracted from any choices about how new technologies may be used. If regulations or other factors can prevent misuse of technology, then faster adoption can become optimal. Finally, we showed that the optimal path of adoption depends on a few parameters, but there is currently a huge amount of uncertainty about the values of these parameters, and careful empirical assessments of the costs and benefits of the adoption of new transformative technologies, such as generative AI, is an obvious area for fruitful research.

References

Acemoglu, D., Aghion, P., Bursztyn, L., & Hemous, D. (2012). The environment and directed technical change. *American economic review*, *102*(1), 131–166.

Acemoglu, D., Aghion, P., & Zilibotti, F. (2006). Distance to frontier, selection, and economic growth. *Journal of the European Economic association*, *4*(1), 37–74.

Acemoglu, D., Antràs, P., & Helpman, E. (2007). Contracts and technology adoption. *American Economic Review*, *97*(3), 916–943.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). Economic policy for artificial intelligence. *Innovation policy and the economy*, *19*(1), 139–159.

Aschenbrenner, L. (2020). *Existential risk and growth* (tech. rep.). GPI Working Paper.

Barro, R. J. (2006). Rare disasters and asset markets in the twentieth century. *The Quarterly Journal of Economics*, *121*(3), 823–866.

Barro, R. J. (2009). Rare disasters, asset prices, and welfare costs. *American Economic Review*, *99*(1), 243–264.

Bovenberg, A. L., & Smulders, S. (1995). Environmental quality and pollution-augmenting technological change in a two-sector endogenous growth model. *Journal of public Economics*, *57*(3), 369–391.

Chui, M., Roberts, R., Yee, L., Hazan, E., Singla, A., Smaje, K., Sukharevsky, A., & Zemmel, R. (2023). *The economic potential of generative ai* (tech. rep.). McKinsey & Company.

Comin, D., & Hobijn, B. (2010). An exploration of technology diffusion. *American economic review*, *100*(5), 2031–2059.

Comin, D., & Mestieri, M. (2014). Technology diffusion: Measurement, causes, and consequences. In *Handbook of economic growth* (pp. 565–622). Elsevier.

Foster, A. D., & Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of political Economy*, *103*(6), 1176–1209.

Foster, A. D., & Rosenzweig, M. R. (2010). Microeconomics of technology adoption. *Annu. Rev. Econ.*, *2*(1), 395–424.

Galasso, A., & Luo, H. (2018). Punishing robots: Issues in the economics of tort liability and innovation in artificial intelligence. In *The economics of artificial intelligence: An agenda* (pp. 493–504). University of Chicago Press.

Jones, C. I. (2016). Life and growth. *Journal of political Economy*, *124*(2), 539–578.

Jones, C. I. (2023). The ai dilemma: Growth versus existential risk.

Katz, M. L., & Shapiro, C. (1986). Technology adoption in the presence of network externalities. *Journal of political economy*, *94*(4), 822–841.

Martin, I. W., & Pindyck, R. S. (2015). Averting catastrophes: The strange economics of scylla and charybdis. *American Economic Review*, *105*(10), 2947–2985.

Martin, I. W., & Pindyck, R. S. (2021). Welfare costs of catastrophes: Lost consumption and lost lives. *The Economic Journal*, *131*(634), 946–969.

Parente, S. L., & Prescott, E. C. (1994). Barriers to technology adoption and development. *Journal of political Economy*, *102*(2), 298–321.

Rietz, T. A. (1988). The equity risk premium a solution. *Journal of monetary Economics*, *22*(1), 117–131.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., . . . Dafoe, A. (2023). Model evaluation for extreme risks.

Weitzman, M. L. (2009). On modeling and interpreting the economics of catastrophic climate change. *The review of economics and statistics*, *91*(1), 1–19.

Weitzman, M. L. (2011). Fat-tailed uncertainty in the economics of catastrophic climate change. *Review of Environmental Economics and Policy*.

Online Appendix

Appendix A contains proofs for the results in the main text. In Appendices B, C, and D, we analyze extensions of our benchmark model and discuss the robustness of our main results.

A Main Proofs from the Main Text

In this appendix, we provide proofs for results in the main text.

A.1 Proofs for Section 3

Proof of Proposition 2. Let $L(t) \equiv L(\mu(t), q(t))$ denote the damage threshold at time $t \in (0, \infty)$. Making use of Proposition 1 and the equation $q(t) = q(0) + (g_N - g_O)t$, the damage threshold equals

$$L(t) = \alpha + (\rho - g_N)\left[\frac{\alpha - \exp(-[q(0) + (g_N - g_O)t])}{\mu(t)\lambda\eta} - \frac{\exp(-[q(0) + (g_N - g_O)t])}{\rho - g_O}\right]. \quad \text{(A1)}$$

It is immediate that $L(t)$ is strictly decreasing in g_O, so adoption at time t is nonincreasing in g_O. Note that adoption is also strictly decreasing whenever $L(t) \in \left(\underline{\delta}, \bar{\delta}\right)$.

Considering instead the comparative static with respect to g_N, we can differentiate to find

$$\frac{\partial L(t)}{\partial g_N} = (1 + (\rho - g_N)t)\left[\frac{1}{\mu(t)\lambda\eta} + \frac{1}{\rho - g_O}\right]\exp(-[q(0) + (g_N - g_O)t]) - \frac{\alpha}{\mu(t)\lambda\eta}.$$

This derivative is positive iff

$$(1 + (\rho - g_N)t)\left[1 + \frac{\mu(t)\lambda\eta}{\rho - g_O}\right]\exp(-[z(0) + (g_N - g_O)t]) \geq \alpha.$$

The left side limits to zero as $t \to \infty$, so there exists an earliest time $\bar{t} < \infty$ such that $\partial L(t)/\partial g_N < 0$ for $t > \bar{t}$. If $\bar{t} > 0$, then the left side of the inequality above must be decreasing in t at $t = \bar{t}$. Since the left side is also decreasing in g_N, we must have that $\bar{t}$ is decreasing in g_N.

Finally, the bracketed term in (A1) limits to a finite value as g_N increases to ρ, which implies that $L(t)$ limits to α. Since the lower support of F satisfies $\alpha \leq \underline{\delta}$, we conclude that $X(\mu(t), q(t))$ limits to zero. ■

Proof of Proposition 3. Using the expression for the damage threshold (7), we can calculate

$$\frac{\dot{L}(\mu,q)}{\rho-g_N} = \frac{1-\mu}{\mu}\frac{\alpha-\exp(-q)}{\eta} + \left(\frac{1}{\mu\lambda\eta}+\frac{1}{\rho-g_O}\right)(g_N-g_O)\exp(-q).$$

This equation implies that $\dot{L}(\mu,q)$ is strictly decreasing in g_O. Differentiating implies that $\dot{L}(\mu,q)$ is strictly decreasing in g_N iff

$$\alpha\exp(q)-1 > \frac{(\rho-g_N)-(g_N-g_O)}{1-\mu}\left(\frac{1}{\lambda}+\frac{\mu\eta}{\rho-g_O}\right).$$

Differentiating $\dot{L}(\mu,q)$ again yields

$$\frac{\ddot{L}(\mu,q)}{\rho-g_N} = \lambda\frac{1-\mu}{\mu}\frac{\alpha-\exp(-q)}{\eta} + \left[2\frac{1-\mu}{\mu}-\left(\frac{1}{\mu\lambda}+\frac{\eta}{\rho-g_O}\right)(g_N-g_O)\right](g_N-g_O)\frac{\exp(-q)}{\eta}.$$

Provided that $\alpha > \exp(-q)$, this expression immediately implies that $\ddot{L}(\mu,q) < 0$ iff $g_N-g_O > G(\mu,q)$, where $G(\mu,q)$ is the largest solution to the quadratic equation

$$\lambda\frac{1-\mu}{\mu}\frac{\alpha-\exp(-q)}{\eta} + \left[2\frac{1-\mu}{\mu}-\left(\frac{1}{\mu\lambda}+\frac{\eta}{\rho-g_O}\right)G(\mu,q)\right]G(\mu,q)\frac{\exp(-q)}{\eta} = 0.$$

Equivalently,

$$G(\mu,q) = \lambda\frac{1+\sqrt{1+\left(1+\frac{\mu\lambda\eta}{\rho-g_O}\right)(1-\mu)^{-1}(\alpha\exp(q)-1)}}{\left(1+\frac{\mu\lambda\eta}{\rho-g_O}\right)(1-\mu)^{-1}}.$$

We observe that G is decreasing in μ and increasing in q, so that $\dot{G}(\mu,q) > 0$ with

$$\lim_{t\to\infty} G(\mu(t),q(t)) = \infty.$$

■

A.2 Proofs for Section 5

Proof of Proposition 7. With a sector-independent use tax $\tau(\mu,Q)$, it is privately optimal to use technology N before the disaster iff

$$\alpha Q_N - Q_O - \tau(\mu,Q) > \mu\lambda\eta\left[\frac{1}{\rho-g_O}Q_O - \frac{\alpha-\gamma_i}{\rho-g_N}Q_N\right]. \tag{A2}$$

The right side of this inequality is strictly increasing in γ_i. Given an initial state $(\mu(0),Q(0))$, let $\tilde{t}_i$ denote the time at which firm i begins using technology N. For any sector j with private damages γ_j, we immediately observe that $\gamma_i \leq \gamma_j$ iff $\tilde{t}_i \leq \tilde{t}_j$. The latter inequality is strict if $\gamma_i < \gamma_j$ and $\tilde{t}_j > 0$.

If γ_i and δ_i are positively affiliated, the tax (12) suffices to implement socially optimal technology choices in equilibrium. To see this, note that the private optimality condition (A2) implies that firm i uses technology N in state (μ, Q) iff $\gamma_i < \hat{L}(\mu, q)$, where

$$\frac{\hat{L}(\mu,q) - \alpha + L(\mu,q) - \kappa(L(\mu,q))}{\rho - g_N} = \frac{\alpha - \exp(-q)}{\mu\lambda\eta} - \frac{\exp -q}{\rho - g_O}.$$

Using the definition of the damage threshold $L(\mu,q)$ from Proposition 1, this equation reduces to $\hat{L}(\mu,q) = \kappa(L(\mu,q))$. Since κ is strictly increasing, we conclude that equilibrium technology choices are efficient: Firm i uses technology N iff

$$\delta_i = \kappa^{-1}(\gamma_i) < \kappa^{-1}\left(\hat{L}(\mu,q)\right) = L(\mu,q).$$

Finally, fix an initial state $(\mu(0), Q(0))$ such that $L(\mu(0), Q(0)) < \underline{\delta}$, so that it is inefficient for any sector to use technology N at $t = 0$. Suppose that a given sector-independent tax $\tau(\mu, Q)$ implements socially optimal technology choices in equilibrium. We can define the affiliation function κ as follows: For any value of social damages $\delta \in [\underline{\delta}, \bar{\delta}]$, let $t(\delta) > 0$ be the time at which sectors with social damages δ (socially) optimally begin using technology N. Since $\tau(\mu, Q)$ implements socially optimal technology choices in equilibrium, these same sectors must find it privately optimal to begin using technology N at time $t(\delta)$. These sectors must have a common value of private damages $\gamma(t(\delta))$: If one sector had a larger value of private damages $\gamma' > \gamma(t(\delta))$, it would find it privately optimal to delay using technology N, contradicting the assumption that τ implements socially optimal technology choices. As a result, the affiliation function $\kappa(\delta) = \gamma(t(\delta))$ is well-defined, and we conclude that social and private damages must be positively affiliated.

■

Proof of Proposition 8. Given a threshold $\hat{\delta}$ and wait time $\hat{T}$, the planner's objective discounted to $t = 0$ can be written

$$\begin{aligned} V\left(\hat{\delta},\hat{T}\right) = &\int_0^{\hat{T}} \exp\left(-\rho t\right) \int_{\delta_i<\hat{\delta}} \Big\{\left(1 - x\left(\mu(t), q(t), \gamma_i\right)\right)\left[1 + \mu(t)\lambda\eta\frac{1}{\rho - g_O}\right] Q_O(t) \\ &\qquad + x\left(\mu(t), q(t), \gamma_i\right)\left[\alpha + \mu(t)\lambda\eta\frac{\alpha - \delta_i}{\rho - g_N}\right] Q_N(t)\Big\}\, didt \\ &+ \int_0^{\hat{T}} \exp\left(-\rho t\right) \int_{\delta_i\geq\hat{\delta}} \left[1 + \mu(t)\lambda\eta\frac{1}{\rho - g_O}\right] Q_O(t) didt \\ &+ \int_{\hat{T}}^{\infty} \exp\left(-\rho t\right) \int_0^1 \Big\{\left(1 - x\left(\mu(t), q(t), \gamma_i\right)\right)\left[1 + \mu(t)\lambda\eta\frac{1}{\rho - g_O}\right] Q_O(t) \\ &\qquad + x\left(\mu(t), q(t), \gamma_i\right)\left[\alpha + \mu(t)\lambda\eta\frac{\alpha - \delta_i}{\rho - g_N}\right] Q_N(t)\Big\}\, didt. \end{aligned}$$

Here $x\left(\mu, q, \gamma_i\right)$ denotes the unrestricted equilibrium technology choice given state (μ, q) and private damages γ_i:

$$x\left(\mu, q, \gamma_i\right) = \begin{cases} 1 & \text{if } \alpha_i - \exp(-q) > \mu\lambda\eta\left[\frac{1}{\rho-g_O}\exp(-q) - \frac{\alpha-\gamma_i}{\rho-g_N}\right], \\ 0 & \text{else.} \end{cases}$$

With $\hat{\delta}$ fixed, we can differentiate V with respect to $\hat{T}$ to find

$$\exp\left(\rho\hat{T}\right)\frac{\partial V\left(\hat{\delta},\hat{T}\right)}{\partial\hat{T}} = -\int_{\delta_i\geq\hat{\delta}} x\left(\mu, q, \gamma_i\right)\left\{\alpha Q_N - Q_O - \mu\lambda\eta\left[\frac{1}{\rho - g_O}Q_O - \frac{\alpha - \delta_i}{\rho - g_N}Q_N\right]\right\} di.$$

To simplify notation, we have left the dependence of the state (μ, Q) on the wait time $\hat{T}$ implicit. First observe that the optimal wait time is bounded:

$$\lim_{\hat{T}\to\infty}\frac{1}{Q_N\left(\hat{T}\right)}\frac{\partial V\left(\hat{\delta},\hat{T}\right)}{\partial\hat{T}} = -\alpha\int_{\delta_i\geq\hat{\delta}} di < 0.$$

Let $\tilde{t}_i$ denote the equilibrium time of adoption for sector i when unrestricted, and let $\underline{t}(\hat{\delta}) \geq 0$ denote the greatest lower bound for these times across all sectors above the threshold ($\delta_i \geq \hat{\delta}$). Note that we can write

$$\exp\left(\rho\hat{T}\right)\frac{\partial V\left(\hat{\delta},\hat{T}\right)}{\partial\hat{T}} = -\int_{\delta_i\geq\hat{\delta},\tilde{t}_i\leq\hat{T}} \alpha Q_N - Q_O - \mu\lambda\eta\left[\frac{1}{\rho - g_O}Q_O - \frac{\alpha - \delta_i}{\rho - g_N}Q_N\right] di.$$

Clearly $\partial V(\hat{\delta})/\partial \hat{T} = 0$, and $\partial V(\hat{\delta})/\partial \hat{T} > 0$ for $\hat{T}$ in a neighborhood of $\underline{t}\left(\hat{\delta}\right)$, because

$$\frac{\partial}{\partial \hat{T}} \exp\left(\rho \hat{T}\right) \frac{\partial V\left(\hat{\delta}, \hat{T}\right)}{\partial \hat{T}}\Bigg|_{\hat{T}=\underline{t}(\hat{\delta})} = -\int_{\delta_i \geq \hat{\delta}, \tilde{t}_i = \underline{t}(\hat{\delta})} \alpha Q_N - Q_O - \mu \lambda \eta \left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \delta_i}{\rho - g_N} Q_N\right] di.$$

On the right-hand side, the state (μ, Q) is evaluated at $\underline{t}(\hat{\delta})$. Since $\gamma_i < \delta_i$ for all sectors above the threshold, the right-hand side must be strictly positive. This implies that V is strictly increasing in $\hat{T}$ in a neighborhood of $\underline{t}(\hat{\delta})$, so the optimal wait time $\hat{T}$ must be interior. It satisfies the first-order condition

$$0 = -\int_{\delta_i \geq \hat{\delta}} x\left(\mu, q, \gamma_i\right) \left\{\alpha Q_N - Q_O - \mu \lambda \eta \left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \delta_i}{\rho - g_N} Q_N\right]\right\} di.$$

Setting $\hat{T} = \underline{t}(\hat{\delta})$ replicates the laissez-faire equilibrium, so this argument establishes that a sandbox policy with $\hat{\delta} > \underline{\delta}$ can strictly improve upon the laissez-faire equilibrium. ■

B Analysis of the General Model

In this part of the Appendix, we analyze the benchmark model without restrictions on the parameters α_i, δ_i, γ_i.

B.1 Socially Optimal Technology Choice

As described in the main text, the planner uses technology N after the disaster iff $\bar{x}_i = 1$ or $\left(\alpha_i - \gamma_i\right) Q_N > Q_O$. Letting $q = \log\left(Q_N / Q_O\right)$ denote the log quality gap between the technologies, we can equivalently define a threshold gap q_i such that the planner uses technology N after the disaster iff $\bar{x}_i = 1$ or $q \geq q_i$:

$$q_i = \begin{cases} -\log\left(\alpha_i - \delta_i\right) & \text{if } \alpha_i > \delta_i, \\ \infty & \text{else.} \end{cases} \tag{B1}$$

At the onset of the disaster, if $q < q_i$ the planner optimally reverts to using technology O in sector i if possible. If $q_i < \infty$, the planner eventually uses technology N again when it attains a sufficiently large lead over technology O.

With this characterization, we can directly integrate the post-disaster HJB equation (4) and take expectations with respect to $\bar{x}$:

$$\begin{aligned} \mathbb{E}\left[W(\bar{x}, Q) \mid x\right] = \int_0^1 \left(1 - x_i \eta_i\right) \Bigg\{ & \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}\left(q_i - q\right)_+\right)\right] \frac{1}{\rho - g_O} Q_O \\ & + \exp\left(-\frac{\rho - g_N}{g_N - g_O}\left(q_i - q\right)_+\right) \frac{\alpha_i - \delta_i}{\rho - g_N} Q_N \Bigg\} + x_i \eta_i \frac{\alpha_i - \delta_i}{\rho - g_N} Q_N \, di. \end{aligned}$$

Here we use the notation $\left(q_i - q\right)_+ = \max\left\{q_i - q, 0\right\}$. Considering the planner's problem before the disaster (3), we observe that it is optimal to use technology N in sector i iff

$$\begin{aligned} \alpha_i Q_N - Q_O > \mu \lambda \eta_i \Bigg\{ & \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}\left(q_i - q\right)_+\right)\right] \frac{1}{\rho - g_O} Q_O \\ & - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}\left(q_i - q\right)_+\right)\right] \frac{\alpha_i - \delta_i}{\rho - g_N} Q_N \Bigg\}. \end{aligned} \tag{B2}$$

This optimality condition differs from (6) because the discounted future net output from using technology O at the time of the disaster now accounts for the possibility that technology N is used after the quality gap q exceeds q_i.

B.2 Comparative Statics for Socially Optimal Adoption

Suppose as in Section 3.2 that α_i and η_i are constant across sectors, but make no assumption about the ranking between δ_i and α. Let $\bar{q}(\delta_i) = q_i$ denote the quality gap above which it is optimal to use technology N in sector i after the disaster (B1), making explicit the dependence on δ_i. The following proposition shows that optimal technology choices can be described using a damage threshold $L(\mu, q)$ and provides comparative statics, generalizing Proposition 1 from Section 3.2.

Proposition B.1. *It is socially optimal to use technology N in sector i before the disaster iff $\delta_i < L(\mu, q)$, where $L(\mu, q)$ is the unique solution to the equation*

$$\alpha - \exp(-q) = \mu\lambda\eta \left\{ \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O} \left(\bar{q}(L(\mu, q)) - q \right)_+ \right) \right] \frac{1}{\rho - g_O} \exp(-q) \right. \qquad \text{(B3)}$$
$$\left. - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O} \left(\bar{q}(L(\mu, q)) - q \right)_+ \right) \right] \frac{\alpha - L(\mu, q)}{\rho - g_N} \right\}.$$

$L(\mu, q)$ (and thus $X(\mu, q)$) is strictly increasing in α and q and strictly decreasing in g_O, λ, and μ. It is strictly decreasing in g_N if $L(\mu, q) > \alpha$ and strictly increasing in g_N if $L(\mu, q) < \alpha$.

Proof. Throughout the proof, we suppress the arguments of the damage threshold $L(\mu, q)$ to simplify notation. The results described in the proposition are easier to prove if we re-write the discounted values on the right-hand side of (B3) as integrals over time. To do this, given a quality gap q, let $\bar{T}(q, g, \delta)$ denote the length of time after the disaster during which it is optimal to use technology O instead of technology N in a sector with damages δ:

$$\bar{T}(q, g, \delta) = \begin{cases} \max\left\{ \frac{-\log(\alpha - \delta) - q}{g_N - g_O}, 0 \right\} & \text{if } \alpha > \delta, \\ \infty & \text{else.} \end{cases}$$

If the sector is not constrained to technology N, its discounted net output after the disaster is

$$\int_0^{\bar{T}(q,g,\delta)} \exp(-\rho t) \exp(g_O t) Q_O dt + \int_{\bar{T}(q,g,\delta)}^{\infty} \exp(-\rho t) \exp(g_N t)(\alpha - \delta) Q_N dt. \qquad \text{(B4)}$$

Similarly, its discounted net output when constrained to technology N is

$$\int_0^{\infty} \exp(-\rho t) \exp(g_N t)(\alpha - \delta) Q_N dt. \qquad \text{(B5)}$$

The bracketed term in (B3) is the difference between the previous two terms above for the marginal sector (with $\delta = L$), divided by Q_N. The right-hand side of (B3) can then be written

$$\text{RHS} = \mu\lambda\eta \int_0^{\bar{T}(q,g,L)} \exp(-\rho t)\left[\exp(g_O t)\exp(-q) - \exp(g_N t)(\alpha - L)\right] dt.$$

We first demonstrate that, when $\alpha > \exp(-q)$ so that technology N is more productive than technology O, there always exists a unique solution L to (B3). We observe that RHS is continuous in L, equals zero when $L \leq \alpha - \exp(-q)$, and limits to infinity as $L \to \infty$. Moreover, RHS is strictly increasing in L when $L > \alpha - \exp(-q)$: This condition implies $\bar{T}(q, g, L) > 0$, and we can differentiate RHS to find

$$\begin{aligned}
\frac{\partial \text{RHS}}{\partial L} &= \mu\lambda\eta \exp\left(-\rho \bar{T}\right)\left[\exp\left(g_O \bar{T}\right)\exp(-q) - \exp\left(g_N \bar{T}\right)(\alpha - L)\right]\frac{\partial \bar{T}}{\partial L} \\
&\quad + \mu\lambda\eta \int_0^{\bar{T}(q,g,L)} \exp(-\rho t)\exp(g_N t)\, dt \\
&= \mu\lambda\eta \int_0^{\bar{T}(q,g,L)} \exp(-\rho t)\exp(g_N t)\, dt \\
&> 0.
\end{aligned}$$

Note that the second equality holds by the Envelope Theorem: $\bar{T}$ maximizes the discounted net output from the marginal sector after the disaster, assuming its technology choice is unconstrained. As a result RHS does not vary locally with respect to $\bar{T}$ ($\partial \text{RHS}/\partial \bar{T} = 0$). Given these properties of RHS, the Intermediate Value Theorem guarantees a unique solution L to (B3) when $\alpha > \exp(-q)$. Moreover, it follows from the optimality condition (B2) that it is socially optimal to use technology N in sector i before the disaster iff $\delta_i < L(\mu, q)$.

The comparative statics for the damage threshold L follow from the Implicit Function Theorem. Holding L fixed, we immediately observe that RHS is decreasing in α and increasing in μ, λ, and η. Differentiating with respect to q, g_O, and g_N yields

$$\begin{aligned}
\frac{\partial \text{RHS}}{\partial q} &= -\mu\lambda\eta \int_0^{\bar{T}} \exp(-\rho t)\exp(g_O t)\exp(-q)\, dt, \\
\frac{\partial \text{RHS}}{\partial g_O} &= \mu\lambda\eta \int_0^{\bar{T}} \exp(-\rho t)\exp(g_O t)\exp(-q)\, t dt, \\
\frac{\partial \text{RHS}}{\partial g_N} &= -\mu\lambda\eta \int_0^{\bar{T}} \exp(-\rho t)\exp(g_N t)(\alpha - L(\mu, q))\, dt.
\end{aligned}$$

These expressions imply that RHS is decreasing in q, increasing in g_O, and decreasing (increasing) in g_N iff $\alpha > (<) L(\mu, q)$. Collecting these results, the Implicit Function Theorem delivers the comparative statics stated in the proposition. ■

The proposition demonstrates that almost all comparative statics from Proposition 1 hold without the assumption that social damages always exceed output from technology N after the disaster ($\alpha_i \leq \delta_i$). However, the comparative static with respect to g_N is sensitive to this assumption. When damages in the marginal sector exceed output ($L(\mu,q) > \alpha$), the damage threshold is decreasing in g_N as in Proposition 1. When damages in the marginal sector are below output ($L(\mu,q) < \alpha$), the damage threshold is instead increasing in g_N.

The following proposition generalizes Proposition 2 to provide full comparative statics for adoption with respect to the growth rates g_O and g_N, including both the direct effects described in Proposition B.1 and the indirect effects through the state $(\mu(t), q(t))$.

Proposition B.2. *For all $t > 0$ with $L(\mu(t), q(t)) < \alpha$:*

1. *$X(\mu(t), q(t))$ is decreasing in g_O.*
2. *$X(\mu(t), q(t))$ is increasing in g_N.*
3. *If $q(0)$ is sufficiently low and $X(\mu(t), q(t)) < F(\alpha)$, $X(\mu(t), q(t))$ is bounded strictly below $F(\alpha)$ as g_N approaches ρ, i.e., $\lim_{g_N \uparrow \rho} X(\mu(t), q(t)) < F(\alpha)$.*

The first two results follow from Proposition B.1 after noting that the damage threshold L is increasing in the quality gap q, and in turn the quality gap $q(t)$ at time t is decreasing in g_O and increasing in g_N. The final result of the proposition follows by taking the limit $g_N \uparrow \rho$ in (B3). Notably, in this limit adoption does not tend to either of the extreme values 0 or $F(\alpha)$, in contrast to the corresponding result in Proposition 2. This holds because, for any sector i with $\delta_i < \alpha$, the discounted net output after the disaster tends to infinity as $g_N \uparrow \rho$ regardless of whether the sector is constrained to use technology N after the disaster. However, the *difference* between the discounted net output when unconstrained and the discounted net output when constrained tends to a finite limit. Socially optimal technology choices before the disaster depend on this difference (see B2), so provided that δ_i is sufficiently close to α and the initial quality gap $q(0)$ sufficiently low, it can remain optimal to delay using technology N in sector i before the disaster even when $g_N \uparrow \rho$.

We illustrate these results in Figure 3 by depicting adoption curves for a stylized parameterization of the model. We modify the calibration of Figure 1 only by assuming that the distribution of damages δ_i is uniform over $[0,5]$ instead of $[1,5]$. Technology choices for sectors with $\delta_i \in [1,5]$ are exactly as in Section 3, and since these sectors comprise $5/6$ of all sectors in this calibration, the adoption curves in Figure 3 when $X(t) \geq 5/6$ are identical to the adoption curves in Figure 1.

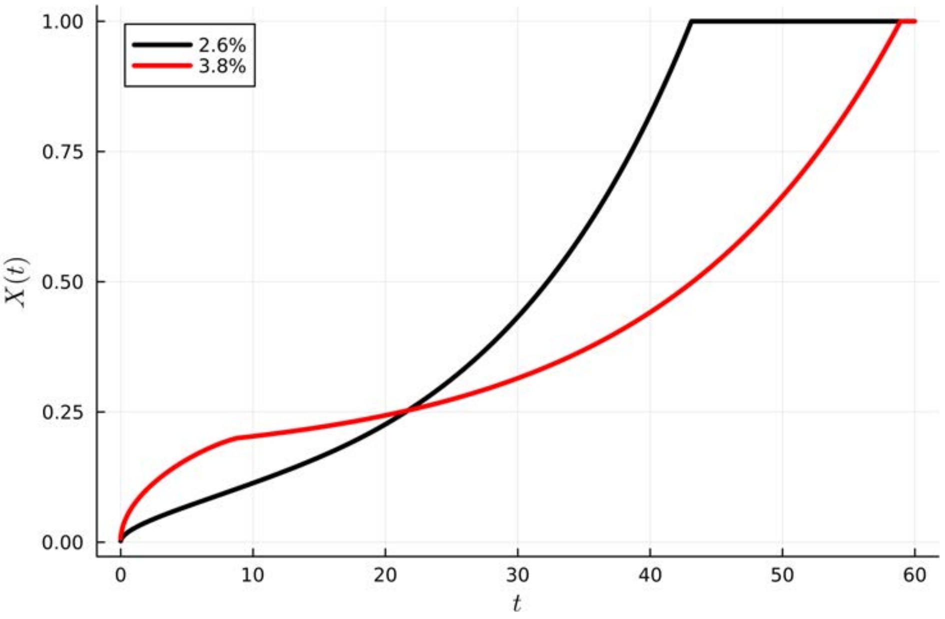

Figure 3: Adoption curves $X(t)$ for different values of g_N. The parameterization is the same as in Figure 1, but with $\underline{\delta} = 0$.

When instead $X(t) \in (0, 1/6)$, the sectors adopting technology N produce positive net output after the disaster, so the analysis in this appendix becomes relevant. In this region, we observe that adoption is increasing in g_N, consistent with Proposition B.2.

B.3 Equilibrium Technology Choice

Using the same derivations as for the optimal technology choice, firm i uses technology N after the disaster iff $\bar{x}_i = 1$ or $q \geq \tilde{q}_i$, where

$$\tilde{q}_i = \begin{cases} -\log\left(\alpha_i - \gamma_i\right) & \text{if } \alpha_i > \gamma_i, \\ \infty & \text{else.} \end{cases}$$

Note that $\tilde{q}_i \leq q_i$ since $\gamma_i \leq \delta_i$. This implies that the private firm returns to using technology N more quickly after the disaster than the planner. Integrating the firm's post-disaster HJB equation (9) and taking expectations with respect to $\bar{x}_i$ yields

$$\begin{aligned}\mathbb{E}\left[\Phi_i\left(\bar{x}_i, Q\right) | x_i\right] = \left(1 - x_i \eta_i\right) \Bigg\{ & \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}\left(\tilde{q}_i - q\right)_+\right)\right] \frac{1}{\rho - g_O} Q_O \\ & + \exp\left(-\frac{\rho - g_N}{g_N - g_O}\left(\tilde{q}_i - q\right)_+\right) \frac{\alpha_i - \gamma_i}{\rho - g_N} Q_N \Bigg\} + x_i \eta_i \frac{\alpha_i - \gamma_i}{\rho - g_N} Q_N.\end{aligned}$$

It is then privately optimal to use technology N in sector i before the disaster iff

$$\alpha_i Q_N - Q_O > \mu\lambda\eta_i \left\{ \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{1}{\rho - g_O} Q_O \right. \\ \left. - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{\alpha_i - \gamma_i}{\rho - g_N} Q_N \right\}. \tag{B6}$$

We observe two differences between this condition and the planner's optimality condition (B2), First, as in the main text, private damages γ_i appear in (B6) instead of the social damages that appear in (B2). Second, the firm begins using technology N more quickly after the disaster than the planner ($\tilde{q}_i \leq q_i$). Both effects tend to reduce the net private cost of irreversibility and incentivize the firm to use technology N more often than the planner before the disaster.

Lemma B.1. *If the social planner uses technology N in sector i in state (μ, Q) before the disaster, then so does firm i.*

Proof. The statement holds provided that firm i's opportunity cost to using technology N instead of technology O at the time of the disaster is smaller than the planner's opportunity cost:

$$\left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{1}{\rho - g_O} Q_O - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{\alpha_i - \gamma_i}{\rho - g_N} Q_N \\ \leq \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}(q_i - q)_+\right)\right] \frac{1}{\rho - g_O} Q_O - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}(q_i - q)_+\right)\right] \frac{\alpha_i - \delta_i}{\rho - g_N} Q_N.$$

Replacing γ_i with δ_i yields the intermediate inequality

$$\left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{1}{\rho - g_O} Q_O - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{\alpha_i - \gamma_i}{\rho - g_N} Q_N \\ \leq \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{1}{\rho - g_O} Q_O - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{\alpha_i - \delta_i}{\rho - g_N} Q_N.$$

Optimality of q_i in the planner's problem after the disaster yields the remaining inequality

$$\left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{1}{\rho - g_O} Q_O - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}(\tilde{q}_i - q)_+\right)\right] \frac{\alpha_i - \delta_i}{\rho - g_N} Q_N \\ \leq \left[1 - \exp\left(-\frac{\rho - g_O}{g_N - g_O}(q_i - q)_+\right)\right] \frac{1}{\rho - g_O} Q_O - \left[1 - \exp\left(-\frac{\rho - g_N}{g_N - g_O}(q_i - q)_+\right)\right] \frac{\alpha_i - \delta_i}{\rho - g_N} Q_N.$$

■

C Extensions

In this part of the Appendix, we discuss two extensions.

C.1 Heterogeneous α_i

Suppose that η_i and δ_i are constant across sectors, and let F_α denote the smooth distribution function for α_i with support $\left[\underline{\alpha}, \bar{\alpha}\right]$. We maintain the assumption that $\alpha_i \leq \delta$ for each sector i, which requires $\bar{\alpha} \leq \delta$. Making use of the planner's optimality condition (6), we observe that there exists a *productivity threshold* $A(\mu, q)$ such that it is optimal to use the new technology in sector i iff $\alpha_i > A(\mu, q)$. Total adoption of the new technology is then the fraction of sectors above the productivity threshold:

$$X(\mu, q) = 1 - F_\alpha(A(\mu, q)).$$

The following proposition characterizes the productivity threshold and is analogous to Proposition 1 in Section 3.2.

Proposition C.1. *It is socially optimal to use technology N in sector i iff $\alpha_i > A(\mu, q)$, where*

$$A(\mu, q) + \mu\lambda\eta \frac{A(\mu, q) - \delta}{\rho - g_N} = \left(1 + \frac{\mu\lambda\eta}{\rho - g_O}\right) \exp(-q). \tag{C1}$$

$A(\mu, q)$ (and thus $1 - X(\mu, q)$) is strictly decreasing in q; strictly increasing in g_O and δ; and strictly increasing in λ, η, μ, and g_N provided that $A(\mu, q) < \delta$.

Proof. The characterizing equation (C1) follows from the planner's optimality condition (6). The comparative statics are immediate from (C1). ■

The analogue of Proposition 2 also holds:

Proposition C.2. *For all $t > 0$:*

1. *$X(\mu(t), q(t))$ is decreasing in g_O.*
2. *There exists an earliest time $\bar{t} < \infty$ such that $X(\mu(t), q(t))$ is decreasing in g_N if $t > \bar{t}$. The time $\bar{t}$ is decreasing in g_N.*
3. *Adoption falls to zero as g_N approaches ρ, i.e., $\lim_{g_N \uparrow \rho} X(\mu(t), q(t)) = 0$.*

Comparative statics for the evolution of the productivity threshold $A(\mu, q)$ over time are less tractable than for the damage threshold $L(\mu, q)$ in the benchmark model. The following proposition provides some guidance about $\dot{A}(\mu,q)$ and $\ddot{A}(\mu,q)$ for the limiting case in which the new and old technologies grow at the same rate.

Proposition C.3. *When $g = g_O = g_N$:*

1. *$\dot{A}(\mu,q)$ is negative and increasing in g.*

2. *There exists a posterior $\bar{\mu} \in (0,1/2)$ such that if $\mu \leq \bar{\mu}$, $\ddot{A}(\mu,q)$ is positive.*

Proof. When $g = g_O = g_N$, the characterizing equation (C1) becomes

$$A(\mu,q) = \frac{1}{1+\frac{\rho-g}{\mu\lambda\eta}}\delta + \exp(-q).$$

The quality gap q is constant since $g = g_O = g_N$. Differentiating in t then yields

$$\begin{aligned}\dot{A}(\mu,q) &= \dot{\mu}\frac{\frac{\rho-g}{\lambda\eta}}{\left(\mu+\frac{\rho-g}{\lambda\eta}\right)^2}\delta,\\ \ddot{A}(\mu,q) &= \left[\ddot{\mu} - 2\dot{\mu}^2\frac{1}{\mu+\frac{\rho-g}{\lambda\eta}}\right]\frac{\frac{\rho-g}{\lambda\eta}}{\left(\mu+\frac{\rho-g}{\lambda\eta}\right)^2}\delta.\end{aligned}$$

Clearly $\dot{A}(\mu,q) < 0$ because $\dot{\mu} < 0$. Using the equations $\dot{\mu} = -\lambda\mu(1-\mu)$ and $\ddot{\mu} = -\lambda\dot{\mu}(1-2\mu)$, we observe that $\ddot{A}(\mu,q) > 0$ iff

$$1-2\mu > 2\frac{\mu(1-\mu)}{\mu+\frac{\rho-g}{\lambda\eta}}.$$

This inequality is violated at $\mu = 1/2$, but it is satisfied at $\mu = 0$. Hence there exists a cutoff $\bar{\mu} \in (0,1/2)$ such that it is satisfied for $\mu \leq \bar{\mu}$. ■

Corollary C.1. *If $g = g_O = g_N$ and $\mu \in (0,\bar{\mu}]$, adoption is concave over time: $\ddot{X}(\mu,q) < 0$.*

These results imply that learning dynamics favor concave adoption over time when sectors are heterogeneous according to comparative advantage, in contrast to the case with heterogeneous damages considered in the main text.

C.2 Constant Damages

In this section, we assess the role of the assumption that post-disaster damages scale with quality Q_N by revisiting the analysis of Section 3 under an alternative assumption: Post-disaster damages in sector i are a fixed constant $\Delta_i \geq 0$. In this case, the planner's HJB equations (3, 4) are still valid, but total damages $D(x)$ are now independent of Q and satisfy

$$D(x) = \int_0^1 x_i \Delta_i di.$$

The planner uses technology N in sector i after the disaster iff $\bar{x}_i = 1$ or $\alpha_i Q_N - \Delta_i > Q_O$. If the disaster strikes when the quality vector is Q and the technology choice in sector i is unconstrained, the planner uses technology O for a time period of length $\bar{T}(Q, g, \Delta_i)$, after which she switches to technology N. The time period $\bar{T}(Q, g, \Delta_i)$ is equal to zero if $\alpha_i Q_N - \Delta_i \geq Q_O$, and otherwise it is the unique solution to the equation

$$\alpha_i Q_N \exp\left(g_N \bar{T}(Q, g, \delta_i)\right) - \Delta_i = Q_O \exp\left(g_O \bar{T}(Q, g, \delta_i)\right).$$

The solution always exists and is unique since $g_N > g_O$.

By the same argument as in Section 3.1, technology N is used in sector i before the disaster if the increase in flow output $\alpha_i Q_N - Q_O$ dominates the expected loss due to the disaster. The latter is the product of the expected arrival rate of the disaster $\mu\lambda$, the probability of irreversibility η_i, and the difference between the discounted value of net output when technology choice is unconstrained and when it is constrained to technology N. If the technology choice in sector i is unconstrained after the disaster, the sector produces discounted net output

$$\int_0^{\bar{T}(Q,g,\delta_i)} \exp(-\rho t) \exp(g_O t) Q_O dt + \int_{\bar{T}(Q,g,\delta_i)}^{\infty} \exp(-\rho t) [\alpha_i \exp(g_N t) Q_N - \Delta_i] dt.$$

When constrained to technology N, the sector's discounted net output is

$$\int_0^{\infty} \exp(-\rho t) [\alpha_i \exp(g_N t) Q_N - \Delta_i] dt.$$

We then that it is optimal to use technology N in sector i before the disaster iff

$$\alpha_i Q_N - Q_O > \mu\lambda\eta_i \int_0^{\bar{T}(Q,g,\delta_i)} \exp(-\rho t) \{\exp(g_O t) Q_O - [\alpha_i \exp(g_N t) Q_N - \Delta_i]\} dt. \qquad \text{(C2)}$$

This optimality condition is analogous to (6) in the benchmark model, but with three differences. First, we have not explicitly integrated the integral in (C2) as we have in (6). Second, in (C2) the fixed damages Δ_i replace the quality-dependent damages $Q_N \delta_i$ in (6). Finally, with quality-independent damages Δ_i it is always optimal to use technology N at some point after

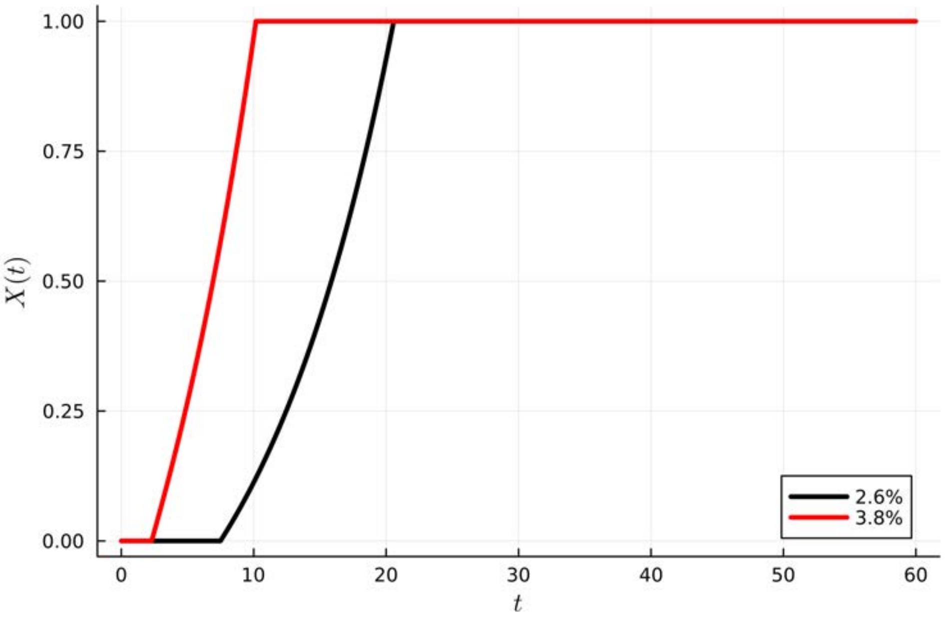

Figure 4: Adoption curves $X(t) \equiv X(\mu(t), q(t))$ for different values of g_N. The parameterization is the same as in Figure 1, but with $\underline{\Delta} = 1$ and $\bar{\Delta} = 5$.

the disaster in sector i: $\bar{T}(Q, g, \delta_i) < \infty$. This contrasts with the benchmark model, in which the assumption $\alpha_i \leq \delta_i$ implies that the planner will always use technology O after the disaster when possible.

Suppose as in Section 3.2 that α_i and η_i are constant across sectors. The following proposition is analogous to Proposition 1 in Section 3.2. It demonstrates that optimal technology choices can be described using a damage threshold $L(\mu, Q)$ and provides comparative statics.

Proposition C.4. *It is socially optimal to use technology N in sector i before the disaster iff $\Delta_i < L(\mu, Q)$, where $L(\mu, Q)$ is the unique solution to the equation*

$$\begin{aligned} &\alpha Q_N - Q_O \\ &\quad = \mu\lambda\eta \int_0^{\bar{T}(Q,g,L(\mu,Q))} \exp(-\rho t)\{\exp(g_O t) Q_O - [\alpha \exp(g_N t) Q_N - L(\mu, Q)]\}\, dt. \end{aligned} \tag{C3}$$

$L(\mu, Q)$ (and thus $X(\mu, Q)$) is strictly increasing in α, Q_N, and g_N and strictly decreasing in g_O, λ, μ, and Q_O.

We omit the proof details, because the argument is almost identical to the proof of Proposition B.1 in Appendix B.2. This proposition demonstrates that, when damages from technology N do not scale with its quality Q_N, optimal adoption is increasing in the growth rate g_N. This contrasts with the corresponding result in Proposition 1, demonstrating that the assumption of proportional damages has significant implications for optimal adoption. We argue that many of the conjectured dangers of (generative) AI more naturally correspond to the case in which damages scale with the capabilities (quality) of rapidly improving models.

We illustrate these results in Figure 4. We modify the calibration of Figure 1 only by assuming constant damages Δ_i uniformly distributed over $\left[\underline{\Delta}, \bar{\Delta}\right]$, where $\underline{\Delta} = 1$ and $\bar{\Delta} = 5$, and by initializing $Q(0) = (1, 1)$. As a result, the initial value of the damages in each sector is the same as in the quantitive example in the main text, as is the quality gap $q(0) = 0$. Consistent with Proposition C.4, we observe that adoption is increasing in the growth rate g_N. Moreover, adoption is much faster than in Figure 1 because (potential) damages do not increase over time as technology N improves.

D Other Issues in Regulation

In this part of the Appendix, we explore second-best tax regulation schemes when private and social damages are not positively affiliated, and we provide additional details about optimal regulatory sandboxes and discuss their advantages relative to sector-independent taxes.

D.1 Second-Best Tax Regulation

Aside from the special case in which social and private damages are positively affiliated, a sector-independent tax cannot implement the optimal technology choices in equilibrium. More generally, use taxes can allow the planner to improve upon laissez-faire technology choices even when optimal ones cannot be implemented. Suppose as in Proposition 7 that α_i and η_i are constant across sectors, but make no assumptions on the joint distribution of δ_i and γ_i. In each state (μ, Q) before the disaster, the planner chooses the use tax $\tau(\mu, Q)$ to maximize output less the expected discounted social cost from the disaster:

$$\max_{\tau} \int_0^1 \left\{ \left(1 - x\left(\mu, Q, \gamma_i, \tau\right)\right) \left[Q_O + \mu\lambda\eta \frac{1}{\rho - g_O} Q_O \right] \right.$$
$$\left. + x\left(\mu, Q, \gamma_i, \tau\right) \left[\alpha Q_N + \mu\lambda\eta \frac{\alpha - \delta_i}{\rho - g_N} Q_N \right] \right\} di.$$

Here $x\left(\mu, Q, \gamma_i, \tau\right)$ describes the equilibrium technology choice for firm i when subject to the tax:

$$x\left(\mu, Q, \gamma_i, \tau\right) = \begin{cases} 1 & \text{if } \alpha Q_N - Q_O - \tau > \mu\lambda\eta \left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \gamma_i}{\rho - g_N} Q_N \right], \\ 0 & \text{else.} \end{cases}$$

Firms adopt technology N in order of increasing γ_i, so we can equivalently assume that the planner selects a private damage threshold $\hat{L}(\mu, q)$ such that firm i uses technology N iff $\gamma_i < \hat{L}(\mu, q)$. The optimal threshold trades off flow consumption against the expected social cost of the disaster. When interior, it satisfies

$$\alpha Q_N - Q_O = \mu\lambda\eta \left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \bar{\delta}\left(\hat{L}(\mu, q)\right)}{\rho - g_N} Q_N \right]. \tag{D1}$$

Here $\bar{\delta}(\gamma) = \mathbb{E}\left[\, \delta_i | \, \gamma_i = \gamma \right]$ is the average social damages across all firms with private damages γ. The optimality condition (D1) is analogous to the original optimality condition (6), but it replaces a single sector's social damages δ_i with the expectation $\bar{\delta}(\gamma)$. The planner's problem is concave iff $\bar{\delta}(\gamma)$ is increasing, in which case an interior solution can be optimal. If, for example,

$\bar{\delta}(\gamma)$ is decreasing, then the planner cannot incentivize sectors with low social damages to use technology N while sectors with high social damages use technology O. As a result, the planner chooses $\hat{L}(\mu, q) = 0$ (no use of N) or $\hat{L}(\mu, q) = \infty$ (full use of N). The latter is optimal when

$$\alpha Q_N - Q_O > \mu\lambda\eta\left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \mathbb{E}[\delta_i]}{\rho - g_N} Q_N\right].$$

D.2 Analysis of Sandbox Regulation

Proposition 8 in the main text demonstrates that it is generally optimal for the planner to implement a regulatory sandbox with a strictly positive wait time $\hat{T}$. The optimal wait time $\hat{T}$ must satisfy the following interior first-order condition, which is derived in the proof of the proposition in Appendix A:

$$0 = -\int_{\delta_i \geq \hat{\delta}} x\left(\mu, q, \gamma_i\right)\left\{\alpha Q_N - Q_O - \mu\lambda\eta\left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \delta_i}{\rho - g_N} Q_N\right]\right\} di. \tag{D2}$$

Here the state (μ, Q) is evaluated at the optimal time $\hat{T}$, and $x\left(\mu, q, \gamma_i\right) = 1$ iff sector i would use technology N in the laissez-faire equilibrium. Two forces determine the optimal wait time $\hat{T}$: If sector i is above the threshold ($\delta_i \geq \hat{\delta}$) and would inefficiently use technology i at time $\hat{T}$, its laissez-faire technology choice would decrease social welfare, favoring a longer wait time:

$$x\left(\mu, q, \gamma_i\right)\left\{\alpha Q_N - Q_O - \mu\lambda\eta\left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \delta_i}{\rho - g_N} Q_N\right]\right\} < 0.$$

If sector i would instead efficiently use technology i at time $\hat{T}$, its laissez-faire technology choice would increase social welfare, favoring a shorter wait time.[15]

We can similarly derive the following interior first-order condition for the optimal threshold $\hat{\delta}$, keeping $\hat{T}$ fixed:

$$0 = \int_0^{\hat{T}} \exp(-\rho t) \int_{\delta_i = \hat{\delta}} x\left(\mu, q, \gamma_i\right)\left\{\alpha Q_N - Q_O - \mu\lambda\eta\left[\frac{1}{\rho - g_O} Q_O - \frac{\alpha - \hat{\delta}}{\rho - g_N} Q_N\right]\right\} dt.$$

If the threshold $\hat{\delta}$ is too high, a large fraction of sectors i face no restrictions on their technology choices, and they subtract too much from social welfare between $t = 0$ and $t = \hat{T}$ as they begin using the new technology too quickly. If $\hat{\delta}$ is too low, then too many sectors i are forced to use technology O between $t = 0$ and $t = \hat{T}$, foregoing the benefits of using technology N in these

[15] As this intuition suggests, it is straightforward to verify that, under the assumptions of Proposition 8, the optimal wait time $\hat{T}$ is nondecreasing in $\hat{\delta}$.

sectors when it is efficient to do so. This analysis demonstrates that the optimal parameters $\left(\hat{\delta}, \hat{T}\right)$ are chosen to resolve a trade-off between restricting early use of the new technology in sectors where expected damages are large, while allowing broad use later as the probability of a disaster falls and the quality gap grows.

We conclude this section by observing that regulatory sandboxes are likely to dominate (or complement) sector-independent taxes when the order of adoption differs substantially between the equilibrium and social optimum. For example, suppose that private and social damages are negatively affiliated: $\gamma_i = \kappa(\delta_i)$, where κ is strictly decreasing. Then Proposition 7 implies that, for any sector-independent tax $\tau(\mu, Q)$, the order in which sectors adopt the new technology in equilibrium is exactly the opposite of the optimal order. Moreover, the analysis in Appendix D.1 implies that the optimal sector-independent tax is such that there exists a time $\hat{T}$ before which no sector uses technology N and after which every sector uses technology N. This time is characterized by the equation

$$\alpha Q_N\left(\hat{T}\right) - Q_O\left(\hat{T}\right) = \mu\left(\hat{T}\right)\lambda\eta\left[\frac{1}{\rho - g_O} Q_O\left(\hat{T}\right) - \frac{\alpha - \mathbb{E}[\delta_i]}{\rho - g_N} Q_N\left(\hat{T}\right)\right].$$

These technology choices can also be implemented using the sandbox policy with threshold $\hat{\delta} = \underline{\delta}$ and wait time $\hat{T}$. Hence a regulatory sandbox can achieve weakly greater social welfare than any sector-independent tax when the misalignment in the order of adoption between the equilibrium and the social optimum is severe.